

Magical Thinking in Data Mining

Charles Elkan (Department of Computer Science and Engineering,
University of California, San Diego)
`elkan@cs.ucsd.edu`

Abstract

CoIL challenge 2000 was a supervised learning contest that attracted 43 entries. The authors of 29 entries later wrote explanations of their work. This paper discusses these reports and reaches three main conclusions. First, naive Bayesian classifiers remain competitive in practice: they were used by both the winning entry and the next best entry. Second, identifying feature interactions correctly is important for maximizing predictive accuracy: this was the difference between the winning classifier and all others. Third and most important, too many researchers do not appreciate properly the issue of statistical significance and the danger of overfitting. Given a dataset such as the one for the CoIL contest, it is pointless to apply a very complicated learning algorithm, or to perform a very time-consuming model search. In either case, one is likely to overfit the training data and to fool oneself in estimating predictive accuracy and in discovering useful correlations.